7-7-2009

A Method for Predicting Citations to the Scientific Publications of Individual Researchers

Peter Z. Revesz

A Method for Predicting Citations to the Scientific Publications of Individual Researchers

Peter Z. Revesz

University of Nebraska-Lincoln
Department of Computer Science and Engineering
Lincoln, NE 68588
1 402 472 3488
revesz@cse.unl.edu

Air Force Office of Scientific Research
875 N. Randolph Street
Arlington, VA 22203
1 703 696 6205
peter.revesz@us.af.mil

ABSTRACT

Any researcher's publications at any time can be ordered from the highest cited to the lowest cited, yielding a citation curve. We describe a novel method for predicting citation curves of researchers in the future. The method depends on treating the citation curves of researchers for various years as one single spatio-temporal function from rank and time to citations. For each researcher, we derive an estimate of this spatio-temporal function that can be used to predict the total citations of individual publications at any given rank at any time. Experiments show that this method can accurately predict entire citation curves and derived measures, such as, the total citations to all publications and the h-index of the researchers.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining, spatial* databases *and GIS*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, search process, selection process.*

General Terms

Algorithms, Measurement, Documentation.

Keywords

Citation, forecasting, h-index, prediction, Web of Science.

1. INTRODUCTION

High citations to a researcher's scientific publications are commonly considered a sign of accomplishment and prominence in his or her field of research. For example, Garfield [4] and Gingras and Wallace [5] found that until the 1960s a significant percentage of the Nobel prizewinners came from the top 500 highest cited authors. Therefore, much data has been compiled to provide citation statistics for individual researchers, for example, the Web of Science database by Thomson Reuters and the online Google Scholar database by Google. Using these databases, one can identify prominent researchers. For example, each year Thomson Reuters identifies the top one percent of the highest cited researchers in several different fields of research and provides their names in a list of Highly Cited Researchers.

However, the problem of citation databases is that researchers are normally past middle age by the time they can be identified as citation leaders in their fields. The same problem plagues other measures of excellence that are based on citations, such as, the h-index measure proposed by J.E. Hirsch [6].

In contrast, in many situations people, such as, members of scientific hiring and promotion committees and grant review panels, are charged with evaluating the research potential of candidates. In that context, it may be beneficial for these people to predict the future citations of candidates as one measure of research potential. Hence we propose a data mining method that uses past citations data to predict future citations. In fact, we predict for any time t in the future and for any individual researcher his or her *citation curve*, which obtained at any time by ordering from highest to lowest-cited the publications of the researcher. For example, Geim's scientific publications at the end of 2012 yield the citation curve shown in Figure 1.

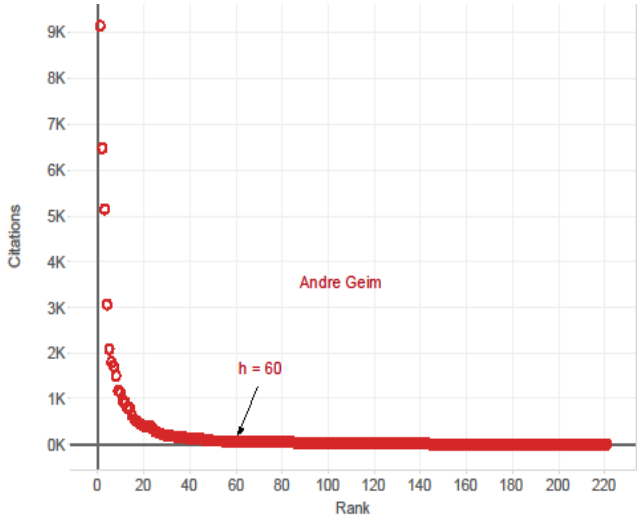


Figure 1. The citation curve of Andre Geim, the 2010 Nobel prizewinner in physics. At the end of 2012, he had 221 publications with a total of 48,414 citations as listed in the Web of Science publications database.

Predicting entire citation curves is a novel task. Previous authors, like Acuna [1] predicted the h-index, which is only a single point of the citation curve, or like Ponomarev et al. [9] and Wang et al. [15] proposed methods for identifying emerging scientific publications, i.e., predicting the long-term impact of single scientific publications. The latter authors commonly look for "breakthrough," "revolutionary," "game changing" or "seminal" publications and start their search by preselecting publications that have a certain minimum number of citations. That number tends to be large and overlook many highly-cited publications. Hence it

IDEAS'14, July 07-09, 2014, Porto, Portugal.

http://dx.doi.org/10.1145/2628194.2628210

does not give a clear picture of most individual researchers' potential.

For example, Andre Geim and Konstantin Novoselov, the 2010 Nobel prizewinners in Physics for their work on graphine have a single publication that has received 9,152 citations by the end of 2012 as listed in the Web of Science database. In fact, Andre Geim had a total of 48,414 citations at that time. However, his total citations was surpassed by Mildred Dresselhaus, a leading carbon chemistry researcher at MIT, and one who is often considered a candidate for the Nobel Prize. At the end of 2012, she had 741 publications, which received a total of 49,102 citations according to the Web of Science database. An approach looking for only breakthrough research may identify Geim and Novosolev's graphine paper but may overlook Dresselhaus' publications. Hence approaches that look for only breakthroughs do not necessarily identify all future scientific leaders, which all three authors clearly are.

A naïve approach to predict the citation curve of an individual researcher A would be to identify the future citations of each publication of A. There are two problems with that approach.

1. First, predicting the future citations of individual publications is very difficult, and many previous attempts were unsuccessful. Even a highly cited publication could be superseded by another publication and lose favor among future authors. Some contributions of other publications become so well-known that many textbooks carry detailed discussions of them, and in this case many authors prefer to cite the textbooks instead of the original publications.
2. Second, even if we could predict perfectly at time t the future citations of the publications that were written by that time by researcher A, we still would not know how many other publications the person will write. Therefore, we cannot predict future citation curves simply by predicting the future of already existing publications. This problem is particularly acute for young researchers, who have most of their publications ahead of them.

In contrast, our approach is to consider the changing and expanding citation curves as *moving or spatio-temporal objects* expanding our research in that area [2,3,7,8,10,11,12,13,14]. We develop an approximation of the spatio-temporal citation curve by a mathematical function from rank and time to citations, where in the approximation the domain of rank and time and the range of citations is the set of rational numbers. Such a generalized spatio-temporal approximation is derived from spatial approximations of the citation curves at a fixed number of already past time instances. In turn, these spatial approximations are complex in themselves because they require several pieces each. Experimentally we found good approximations with only three pieces for most researchers' citation curves, namely, one piece for the top forty percentile of publications, a second piece for the middle twenty percent of publications, and a third piece for the bottom forty percent of publications.

The spatio-temporal citation curve allows predicting at any future time many measures that are based on citation curves, such as, total citations of each publication and the *h-index*, which is the largest rank where the citations are still greater than equal to the rank [6]. Indeed, our spatio-temporal approximation can be used analyze citation curves and answer many complex queries, such as, "What will be the h-index of researcher A at time *t*?" or "When will the h-index of researcher A exceed the h-index of researcher B?" or "When will researcher C have at least ten papers that are each cited over 100 times?"

This paper is organized as follows. Section 2 gives some basic definitions. Section 3 describes our method of approximating citation curves at fixed time instances. Section 4 describes our spatio-temporal citation curve generalization that leads automatically to a prediction method. Section 5 gives experimental results. Finally, Section 6 presents some conclusions and future work.

2. BASIC CONCEPTS

Let the publications of researcher A be A_i for $1 \le i \le n$ for n publications in order of citation rank from highest to lowest cited. Such an ordering yields a *citation curve* from rank to citations, which are both integers. We denote by $c(A_i, t)$ the cumulative total citation count of publication A_i at time t, which is usually the end of some year. We denote by c(A, t) the total citation count for all publications of A_i for $1 \le i \le n$ at time t. Similarly, we denote by $e(A_i, t)$ the estimated cumulative total citation count of publication A_i at time t. We denote by e(A, t) the estimated total citation count for all publications A_i for $1 \le i \le n$ at time t. Finally, we denote by h(A, t) the estimated *h-index* of researcher A at time t. An h-index of researcher A is the number of publications that has greater than equal citations that their ranks.

In this paper, we propose several estimation methods to find $e(A_i, t)$, which gives an estimate of the entire citation curve of researcher A at time t. We are primarily interested in finding estimates with low mean squared error (MSE), defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (e(A_i, t) - c(A_i, t))^2$$

The square root of the MSE is the *standard error*. For the computer experiments in Section 5, the *standard error*, or StdErr, is calculated using the Tableau data analytics and visualization system. Regarding e(A, t) and h(A, t), we are interested in minimizing the percentage difference from c(A, t) and the real h-index value at time t. All material on each page should fit within a rectangle of 18 × 23.5 cm (7" × 9.25"), centered on the page, beginning 1.9 cm (0.75") from the top of the page and ending with 2.54 cm (1") from the bottom. The right and left margins should be 1.9 cm (.75").

3. CITATION CURVES APPROXIMATED BY RATIONAL DOMAIN FUNCTIONS

We piecewise approximate the citation curve of any researcher by functions of the form:

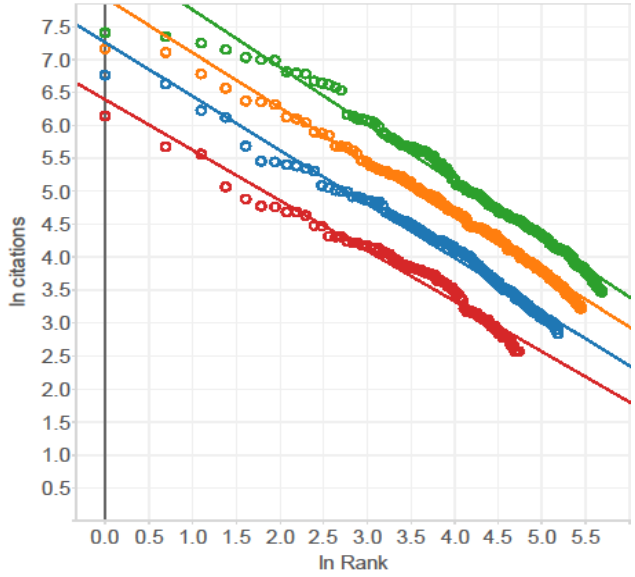
$$y = a\,x^p + b \qquad (1)$$

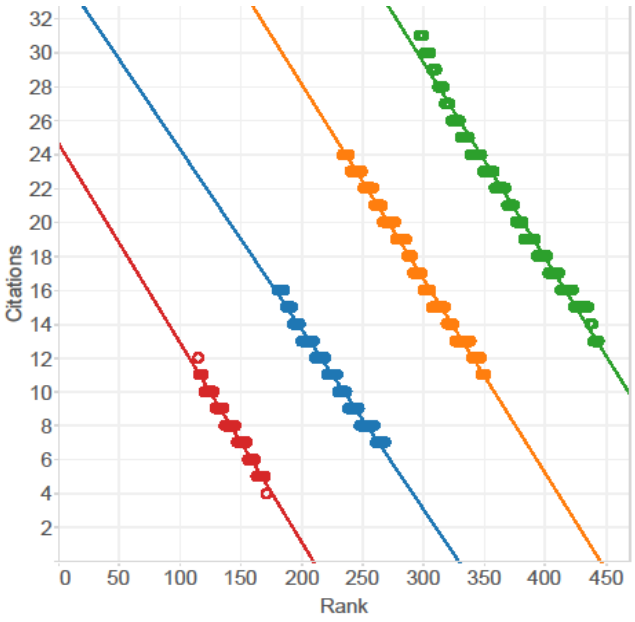
where x is the rank, y is the total citations of the publication of that rank, and a, b and p are rational number constants. While the original citation curve has a discrete domain, the approximation power law function $f(x)$ has a rational domain. We approximate each researcher's citation curve in three pieces using equation (1) with the following restrictions:

Top 40 percentile of publications: $b = 0$, that is, a power law function. The top 40 percentile of Dresselhaus' citation curves in years 2000, 2004 and 2008 can be visualized on a log-log graph as shown in Figure 2A. The near linearity of the curvess shows that there is a power law relationship between rank and citations.

40 to 60 percentile of publications: $p = 1$, that is, a linear function. The linearity of these citation curve pieces is shown in Figure 2B.

Bottom 40 percentile of publications: a = 0, that is, a constant function. Moreover, b = ln(Min_{60}), i.e., the natural logarithm of Min_{60}, which is the total citations to the last publication in the top 60 percentile. Figure 2C shows for the bottom 40 percent of publications, the citations vary with the logarithm of the rank. Active researchers who write many new publications have usually many of those new publications with zero citations. Inactive researchers who stopped publishing new publications have few publications with zero citations. Since researchers can at any time switch from active to inactive, these segments are particularly difficult to predict precisely. However, we found ln(Min_{60}) to be a good approximation.





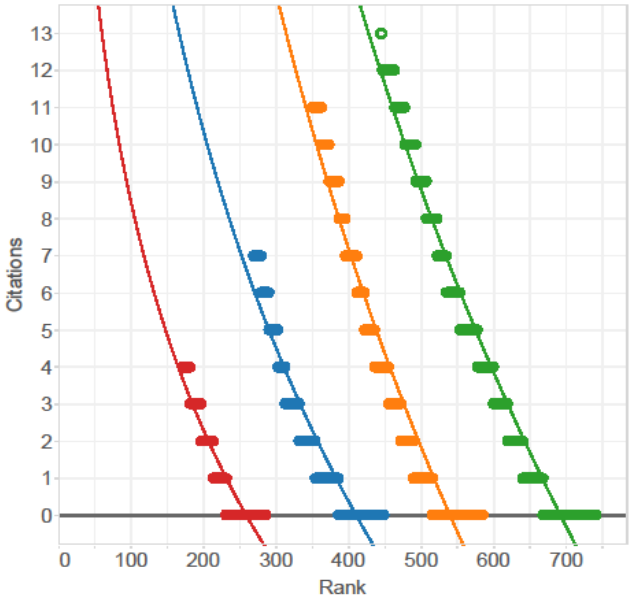


Figure 2. Dresselhaus' citation curves at the end of 2000 (red), 2004 (blue), 2008 (orange) and 2012 (green) are shown for the top 40 percent of the publications (A first picture), for the 40 to 60 percentile of publications (B second picture) and for the bottom 40 percent of publications (C third picture). Note that Figure 2A is displayed as a natural logarithm of rank (x axis) and natural logarithm of citations (y axis) graph.

Example 1. For Dresselhaus' publications at the end of 2000, 2004, 2008 and 2012, we found using the Tableau data analytics and visualization system the best-fit approximations shown in Table 1.

Table 1. Approximations for Dresselhaus' citation curves in 2000, 2004, 2008 and 2012.

Year	Percent	Rank	a	p	b	Min_{60}
2000	0-40	1-114	595.368	0.7645	0	
2004	0-40	1-179	1418.237	0.8176	0	
2008	0-40	1-233	2785.214	0.8318	0	
2012	0-40	1-296	5568.383	0.8715	0	
2000	40-60	115-171	0.117708	1	24.674	
2004	40-60	180-269	0.106194	1	34.941	
2008	40-60	234-351	0.114707	1	51.103	
2012	40-60	297-444	0.115414	1	64.078	
2000	60-100	172-286				4
2004	60-100	270-449				7
2008	60-100	352-585				11
2012	60-100	445-741				13

The functions in Table 1 give close approximations to Dresselhaus' citation curves. Table 2 shows the approximate and the actual citations of some publications of Dresselhaus between 2000 and 2012.

Table 2. Tests of the approximations on some sample publications.

Year	Percent	Rank	Citations Approximate	Citations Actual
2000	0-40	75	21.94	21
2004	0-40	75	41.57	45
2008	0-40	75	76.78	83
2012	0-40	75	129.28	133
2000	40-60	150	7.02	7
2004	40-60	200	13.7	13
2008	40-60	250	22.43	23
2012	40-60	400	17.91	18

4. A CITATION PREDICTION METHOD

Our citation prediction method is based on extending the approximate citation curves $f_0(x)$, $f_1(x)$, $f_2(x)$ ... at fixed times t_0, t_1, t_2 ... respectively, into a spatio-temporal citation curve $\varphi(x, t)$. Figure 3 shows the top 40 percentile of Dresselhaus' citation curves at the end of 2000 (red), at the end of 2004 (blue), at the end of 2008 (orange) and at the end of 2012 (green). These four separate citation curves can be viewed as instances of a spatio-temporal citation curve. Intuitively, in the spatio-temporal citation curve the entire curve expands proportionally. Hence publications at the same percentile correspond to each other. For example, the publications nearest the 10th percentile in each citation curve, that is, publications D_{29} in 2000, D_{45} in 2004, D_{59} in 2008 and D_{74} in 2012 are placed at similar locations on the corresponding citation curves. That correspondence is similar to the correspondence of items that are each one standard deviation above the mean in two normal distribution graphs, one distribution generated from a small sample and the other distribution generated from a large sample of the same population.

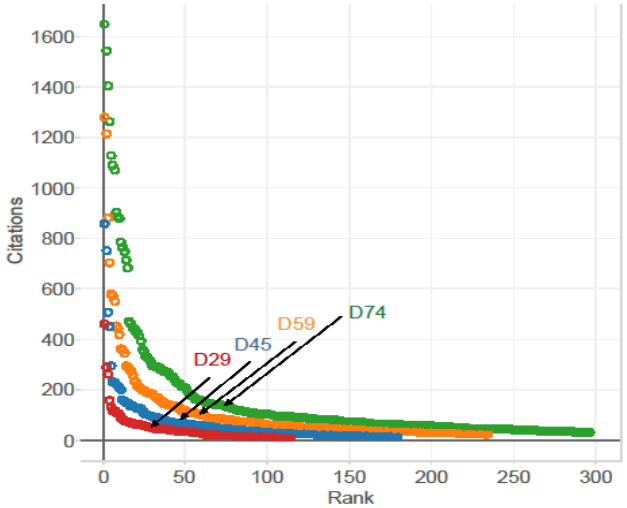


Figure 3. The top 40 percent of Dresselhaus' citation curves at the end of 2000 (red), 2004 (blue), 2008 (orange) and 2012 (green). On each of the curves the publications closest to the 10th percentile, that is, D_{29} in 2000, D_{45} in 2004, D_{59} in 2008 and D_{74} in 2012 are marked by arrows.

The citation prediction method can be described in a theorem as follows.

Theorem 1 For any researcher, let $f_0(x)$, $f_1(x)$ and $f_2(x)$ be the approximation functions for the citation curves and n_0, n_1 and n_2 be the number of publications at times t_0, t_1 and t_2, respectively, where $t_2 > t_1 > t_0$. Then the researcher's spatio-temporal citation curve $\varphi(x, t)$ for any time $t > t_2$ can be estimated to be:

$$\varphi(x,t) = \left(\frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) f_0\left(\frac{n_0 x}{n(t)}\right) - \left(\frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2} + \frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) f_1\left(\frac{n_1 x}{n(t)}\right) + \left(1 + \frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2}\right) f_2\left(\frac{n_2 x}{n(t)}\right) \quad (2)$$

where $n(t)$ is the number of publications at time t. In addition, n(t) could be estimated by:

$$n(t) = \left(\frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) n_0 - \left(\frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2} + \frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) n_1 + \left(1 + \frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2}\right) n_2 \quad (3)$$

Proof: By the proportionality assumption, rank x at time t corresponds to rank n0 x/n, n1 x/n and n2 x/n at times t0, t1 and t2, respectively. For simplicity, let $C_0 = f_0\left(\frac{n_0 x}{n(t)}\right)$, $C_1 = f_1\left(\frac{n_1 x}{n(t)}\right)$ and $C_2 = f_2\left(\frac{n_2 x}{n(t)}\right)$. Between corresponding ranks, the velocity of the citations change at t_1 is:

$$V_1 = \frac{C_1 - C_0}{t_1 - t_0}$$

Similarly, the velocity of the citations change at t2 is:

$$V_2 = \frac{C_2 - C_1}{t_2 - t_1}$$

Therefore the citation change accelerated from t1 to t2 is:

$$A_2 = \frac{V_2 - V_1}{t_2 - t_1} = \frac{C_2 - C_1}{(t_2 - t_1)^2} - \frac{C_1 - C_0}{(t_2 - t_1)(t_1 - t_0)}$$

We can estimate the citations at any time t as follows:

$$\varphi(x,t) = C_2 + (t - t_2)\,V_2 + \frac{(t-t_2)^2}{2}\,A_2$$

Simplifying yields:

$$\varphi(x,t) = \left(\frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) C_0 - \left(\frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2} + \frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) C_1 + \left(1 + \frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2}\right) C_2$$

The above yields Equation (2), and the approximation for n(t) can be obtained similarly. ■

Example 2. Table 1 implies that the top 40 percent of Dresselhaus' citation curves for 2000, 2004 and 2008 can be approximated by the following power law functions:

$$f_{2000,\,0\text{-}40} = 595.368\ x^{-0.7645}$$

$$f_{2004,\,0\text{-}40} = 1418.237\ x^{-0.8176}$$

$$f_{2008,\,0\text{-}40} = 2785.214\ x^{-0.8318}$$

We also know that n2000 = 286, n2004 = 449 and n2008 = 585. Substituting these into Equation (2) yields:

$$\varphi_{0-40}(x,t) = 297.68\left(\frac{286\,x}{n(t)}\right)^{-0.7645} - 2836.47\left(\frac{449\,x}{n(t)}\right)^{-0.8176} + 6963.04\left(\frac{585\,x}{n(t)}\right)^{-0.8318}$$

Similarly, Table 1 implies for the 40-60 percent of Dresselhaus' citation curves the approximations:

$$f_{2000,\,40\text{-}60} = -0.117708\,x + 24.674$$

$$f_{2004,\,40\text{-}60} = -0.106194\,x + 34.941$$

$$f_{2008,\,40\text{-}60} = -0.114707\,x + 51.103$$

Substituting these into Equation (2) yields:

$$\varphi_{40-60}(x,t) = -0.058854\left(\frac{286\,x}{n(t)}\right) + 0.212388\left(\frac{449\,x}{n(t)}\right) - 0.2867675\left(\frac{585\,x}{n(t)}\right) + 70.2127$$

In particular, at exactly the 60 percentile of the approximation graph, we have for any t:

$$\mathrm{Min}_{60} = -0.058854\,(286*0.6) + 0.212388\,(449*0.6) - 0.2867675\,(585*0.6) + 70.2127 \approx 16.7$$

Hence, we estimate:

$$\varphi_{60-100}(x,t) = \ln(16.7)$$

The functions $\varphi_{0-40}(x,t)$, $\varphi_{40-60}(x,t)$ and $\varphi_{60-100}(x,t)$ form the piecewise approximation $\varphi(x,t)$ of the citation curve. ■

We can analyze the spatio-temporal citation curve of any researcher at any time the same way as the actual citation curves are analyzed. For example, we can easily find an estimate of the total citations or the h-index.

Theorem 2 Let any researcher A's spatio-temporal citation curve be *φ(x, t)* and temporal publication function be *n(t)*. Then A's total citations and h-index at time t can be estimated as follows:

$$(A,t) = \sum_{i=1}^{n(t)} \varphi(i,t) \qquad (4)$$

$$h(A,t) = i \quad \text{if } \varphi(i,t) \geq i \text{ and } \varphi(i+1,t) < i+1 \qquad (5)$$

■

Example 3. Suppose that we want to find the total citations and the h-index for Dresselhaus in 2012 using data from 2000, 2004 and 2008. Substituting t_0 = 2000, t_1 = 2004, t_2 = 2008, and t = 2012, n_{2000} = 286, n_{2004} = 449 and n_{2008} = 585 into Equation (3), we estimate the number of publications in 2012 as:

$n_{2012} = 2.5\ n_{2008} - 2\ n_{2004} + 0.5\ n_{2000} \approx 708$

Hence we also assume that the top 40 percent of publications are D_1 to D_{283}, the middle twenty percent are D_{284} to D_{425}, and the bottom twenty percent are D_{426} to D_{708}. Hence substituting into Equation (4) we estimate of the total citations of the publications in the top 40 percent in 2012 as follows:

$$C_{0-40}(A,2012) = \sum_{i=1}^{283} 297.68\left(\frac{286\,i}{708}\right)^{-0.764526} - 2836.47\left(\frac{449\,i}{708}\right)^{-0.817568} + 6963.04\left(\frac{585\,i}{708}\right)^{-0.831757}$$

The above estimate is about 45,929. The actual number of total citations for the top 40 percent of publications was 44,617 in 2012. For the middle 20 percent in 2012, we estimate:

$$C_{40-60}(A,2012) = \sum_{i=284}^{425} -0.058854\left(\frac{286\,i}{708}\right) + 0.212388\left(\frac{449\,i}{708}\right) - 0.2867675\left(\frac{585\,i}{708}\right) + 70.2127$$

The above estimate is 3,626. We note that Min60 = 16.7. Hence for the bottom 40 percent we estimate:

$$C_{60-100}(A, 2012) = \sum_{i=426}^{708} \ln(16.7) \approx 796$$

Hence our citation prediction method estimates that c(A, 2012) = 45,929 + 3,626 + 796 = 50,351 citations. That estimate was within 2.5% of the actual value.

We can use Equation (5) to estimate the h-index too in 2012 by simply calculating in a while loop the estimated citations for each paper until the rank exceeds the total citations for a publication. Figure 4 shows a comparison between the actual and the estimated citation curves of Dresselhaus for 2012. As can be seen the two curves are close to each other. Hence the actual and the estimated h-index values, 101 and 99, respectively, are also close to each other.

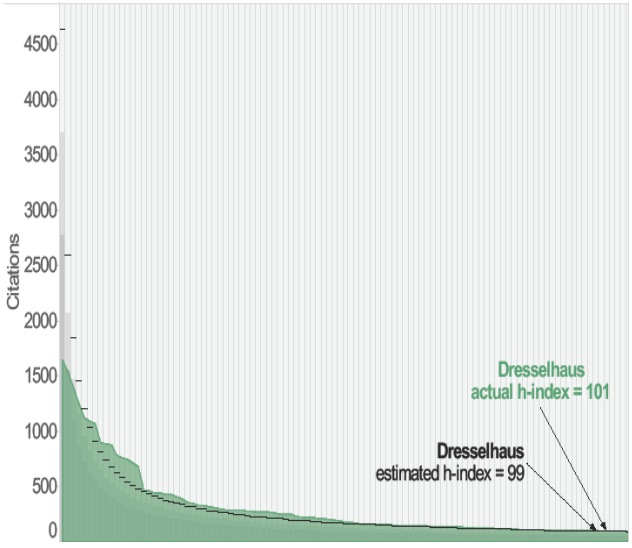


Figure 4. Dresselhaus' actual 2012 citation curve (green area) and estimated 2012 citation curve (black line) for publications ranked from 1 to 105. The estimate was based on citations data from 2000, 2004 and 2008. The close relationship between the two curves allows good estimation of other measures that are based on citations, such as, the h-index. In this case, the actual h-index is 101, while the estimated h-index is 99.

5. EXPERIMENTAL RESULTS

5.1 Sample Data Collection and Statistical Analysis of the Sample

In our experiments we collected citation data from the Web of Science database for eight leading physics researchers. These researchers included Andre Geim and Konstantin Novosolev, winners of the Nobel Prize in Physics in 2010, Serge Haroche and David Wineland, winners of the Nobel Prize in Physics in 2012, Brian P. Schmidt, one of the winners of the Nobel Prize in Physics in 2011, and Mildred Dresselhaus, Geoffrey Marcy, and Didier Queloz. The last three of these researchers are also rumored to be nominees for the Nobel Prize. All of these prominent researchers have a large set of publications, which enabled easier data collection from the Web of Science database and statistically more reliable experiments.

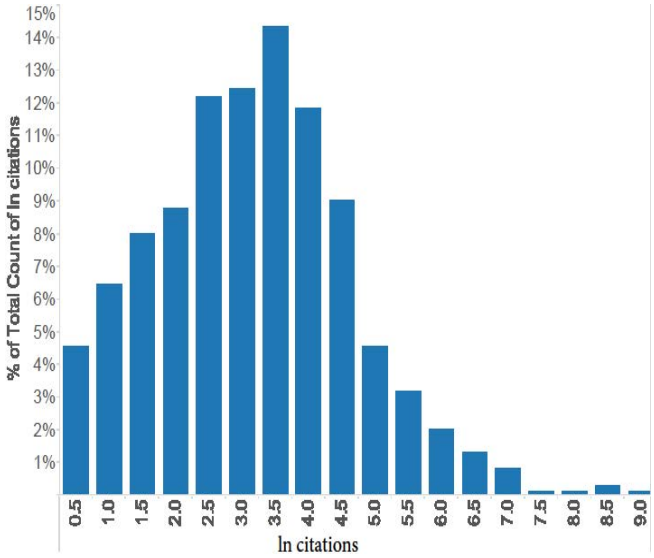


Figure 5. The distribution of the natural logarithms of the citations of the publications of eight physics researchers is approximately a normal distribution. The larger than expected percentage of publications with very low natural logarithm of citations is due mainly to recently published publications that have not received any citations because of their recentness. A more symmetric distribution can be obtained if we only consider publications that were published at least five years ago.

Our first goal was to test the distribution of the citations of these researchers. Figure 5 shows that their citations are approximately log normal distributed. Since other researchers already found log-normal distributions for the citations of large number of researchers, the log-normal distribution of our data set suggests that we have a good representative sample in terms of statistical distribution even though even each of the selected researchers had a larger number of publications than usual. Previously, one of our concerns was that the Web of Science database may ignore many low impact journals, which would result in a lower than expected number of uncited publications. In contrast, the distribution in Figure 5 deviated from a perfect log-normal distribution in a somewhat larger number of uncited or rarely cited publications. That deviation which was probably due to the fact that all of the researchers were still active, that is, they continued to publish many new publications that had little chance yet to be read and cited by other researchers.

5.2 Experiments Testing the Accuracy of the Piecewise Approximation Method for Citation curves

Next we mapped all the researchers' top 40 percentile citation curves onto a log-log graph, except for Dresselhaus, whose citation curve we displayed already in Figures 2 and 3. Figure 6 shows that for the seven other seven researchers also the top 40 percentile tended to have a power law relationship between rank and citations of publications at the end of 2012. Figure 7 shows a linear relationship for the middle twenty percentile. Hence the trends noticed for Dresselhaus' publications hold as well for the other researchers.

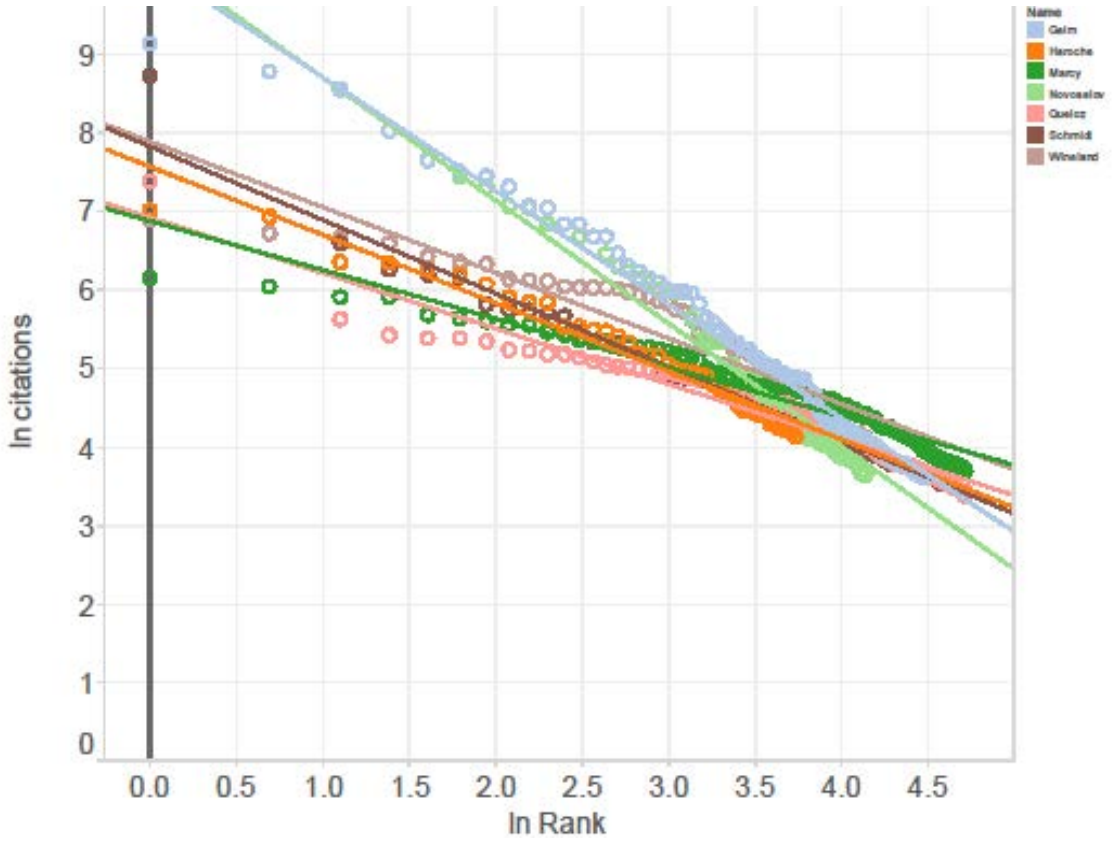


Figure 6. A power law between rank and citations describes the top 40 percentile of publications of the seven other leading physics researchers in 2012.

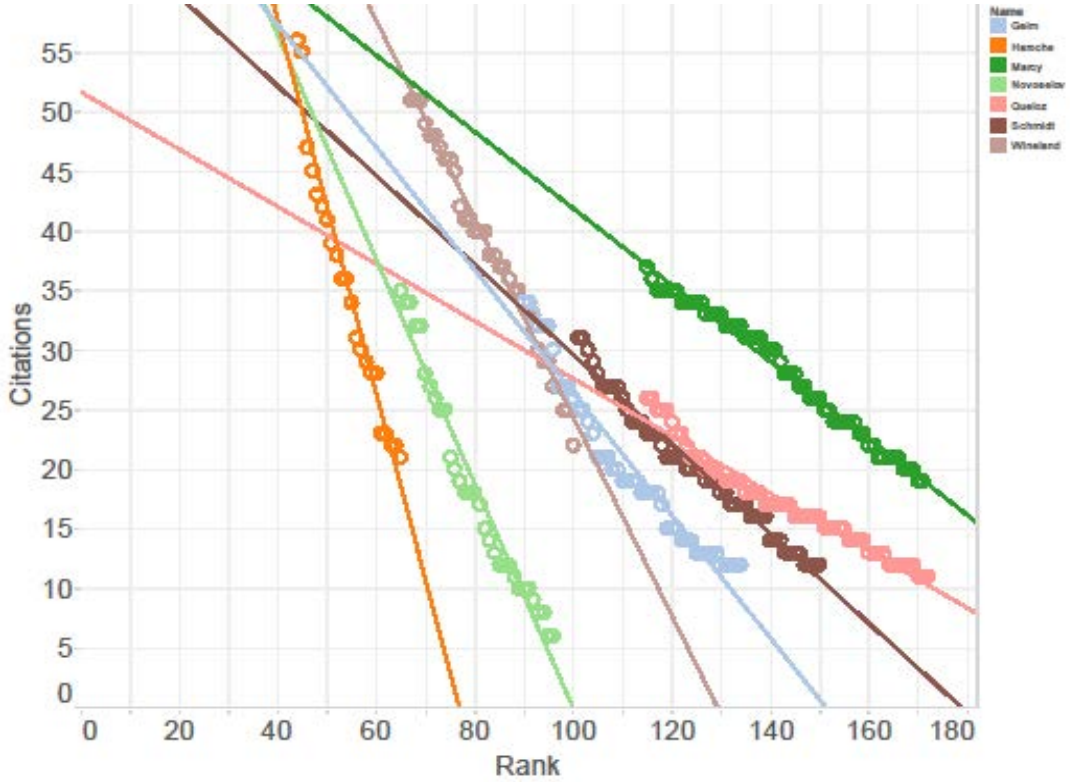


Figure 7. A linear function between rank and citations describes the 40-60 percentile of publications of the seven other leading physics researchers in 2012.

Table 3. The experimental results for the piecewise approximations of the citation curves for 2012.

Name	Percent	Rank	a	p	b	Min_{60}	StdErr	Sum Estimate	Sum Actual	Error %
Dresselhaus	0-40	1-296	5568.383	-0.871541	0		0.00717	49871	44617	
Dresselhaus	40-60	297-444	-0.115414	1	64.0783		0.00103	3151	3155	
Dresselhaus	60-100	445-741				13		762	1330	
Dresselhaus	0-100							53784	49102	9.5
Geim	0-40	1-88	25714.235	-1.44745	0		0.02101	65392	47198	
Geim	40-60	89-132	-0.54179	1	80.8223		0.01879	922	922	
Geim	60-100	133-221				12		221	294	
Geim	0-100							66535	48414	37
Haroche	0-40	1-42	1938.675	-0.8687	0		0.02921	10492	9917	
Haroche	40-60	43-64	-1.65895	1	125.39		0.07366	806	806	
Haroche	60-100	65-107				22		133	309	
Haroche	0-100							11431	11032	3.6
Marcy	0-40	1-112	968.7	-0.621212	0		0.01656	13276	12750	
Marcy	40-60	113-168	-0.321623	1	74.0623		0.00483	1617	1616	
Marcy	60-100	169-281				20		339	765	
Marcy	0-100							15232	15131	0.7
Novosolev	0-40	1-63	28787.6	-1.56564	0		0.03665	63787	42834	
Novoselov	40-60	64-95	-0.967925	1	96.1688		0.03419	615	615	
Novosolev	60-100	96-159				6		115	101	
Novosolev	0-100							64517	43550	48
Queloz	0-40	1-112	1013.292	-0.707164	0		0.01689	10899	10931	
Queloz	40-60	113-169	-0.250972	1	53.1941		0.00745	1015	1015	
Queloz	60-100	170-282				12		281	387	
Queloz	0-100							12195	12333	1.1
Schmidt	0-40	1-98	2502.486	-0.935619	0		0.01234	14797	17519	
Schmidt	40-60	99-148	--0.39078	1	69.1014		0.00808	1042	1042	
Schmidt	60-100	149-247				12		246	327	
Schmidt	0-100							16085	18888	14.8
Wineland	0-40	1-65	2667.268	-0.835968	0		0.03906	17536	16185	
Wineland	40-60	66-99	-0.826738	1	107.235		0.01457	1327	1327	
Wineland	60-100	100-165				25		212	438	
Wineland	0-100							19075	17950	6.3

Table 4. The experimental results for the piecewise approximations for 2000, 2004 and 2008 (blue lines) and the predictions for the total citations and the h-index for 2012 (red lines).

Name	Year	Percent	Rank	a	p	b	Min_{60}	StrErr	Sum Est.	Sum Act.	Error %
Dresselhaus	2000	0-40	1-114	595.368	-0.764526	0			5526	5208	
Dresselhaus	2000	40-60	115-171	-0.117708	1	24.67			447	447	
Dresselhaus	2000	60-100	172-286				4		159	138	
Dresselhaus	2000	0-100	1-286						6132	5793	5.85
Dresselhaus	2004	0-40	1-179	1418.237	-0.817568	0			13064	12173	
Dresselhaus	2004	40-60	180-269	-0.106194	1	34.94			999	999	
Dresselhaus	2004	60-100	270-449				7		350	373	
Dresselhaus	2004	0-100	1-449						14413	13545	6.4
Dresselhaus	2008	0-40	1-233	2785.214	-0.831757	0			26454	24253	
Dresselhaus	2008	40-60	234-351	-0.114707	1	51.1			2070	2071	
Dresselhaus	2008	60-100	352-585				11		561	857	
Dresselhaus	2008	0-100	1-585						29085	27181	7
D. citations	2012	0-100	1-708						50351	49102	2.5
D. h-index	2012								99	101	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Wineland	2000	0-40	1-30	530.28	-0.760093	0		0.03371	3105	3029	
Wineland	2000	40-60	31-45	-1.07857	1	62.85		0.06203	328	328	
Wineland	2000	60-100	46-76				15		84	104	
Wineland	2000	0-100	1-76						3517	3461	1.6
Wineland	2004	0-40	1-41	1130.45	-0.799404	0		0.02847	6899	6585	
Wineland	2004	40-60	42-62	-1.22468	1	94.64		0.05363	650	650	
Wineland	2004	60-100	63-104				18		121	208	
Wineland	2004	0-100	1-104						7670	7443	3
Wineland	2008	0-40	1-51	1693.83	-0.787413	0		0.03768	11401	10795	
Wineland	2008	40-60	52-76	-1.17	1	111.36		0.0493	912	912	
Wineland	2008	60-100	77-128				24		165	381	
Wineland	2008	0-100	1-128						12478	12088	3.2
W. citations	2012	0-100	1-150						17833	17950	0.7
W. h-index	2012								60	61	1.6

Table 3 shows the details of the piecewise linear approximations. In all cases the standard error was below 0.0736611, and the p-value was below 0.0001. For most researchers the difference between the estimated and the actual total citations were less than 15%. The exceptions were Geim and Novoselov, who coauthored on graphine a celebrated and highly cited publication, which accounts for most of the differences between their estimated and actual total citations.

5.3 Experiments Testing the Accuracy of Predicting Total Citations and H--Index

We also tested the accuracy of predicting the total citations and the h-index in 2012 using data from 2000, 2004, 2008 and 2012. The results are displayed in Table 4. In general the higher accuracy approximations for 2000, 2004, 2008 also yielded higher accuracy estimates for 2012. For example, for Wineland, where the approximations were all within 3.2 percent, the 2012 estimate was within 0.7 percent for total citations and 1.6 for the h-index.

6. CONCLUSIONS AND FURTHER WORK

We gave a prediction method for citations to all the publications of an individual researcher. We also experimented with a small set of physics researchers. The experiments show that the citation prediction method gave estimates that were close to the actual data. The experiments focused on Nobel Prize winners because their have extensive and well-documented publication records. The method's good performance on the early career years of these celebrated scientists, that is, when they had only a modest number of publications and citations, suggests that the method could work also well for ordinary researchers.

In the future, we would like to extend the experiments to a larger set of researchers and to include researchers in other scientific areas, including computer science, biology and chemistry. We also plan to experiment with approximations that are more refined, i.e., consist of more than just three pieces. Such approximations could further enhance the accuracy of the prediction method. We would also like to make experiments that test how accuracy changes with the number of years by which we try to predict ahead the total citations and the h-index.

7. ACKNOWLEDGMENTS AND DISCLAIMER

This research was done while the author was an AAAS Science and Technology Policy Fellow. The views expressed in this paper are those of the author and do not necessarily reflect the views of the US federal government or its agencies.

8. REFERENCES

[1] Acuna, D. E., Allesina, S., and Kording. K. P. 2012. Future impact: Predicting scientific success. *Nature*, 489, 7415 (September 2012), 201-202.

[2] Anderson, S. and Revesz, P. Z. 2009. Efficient MaxCount and threshold operators of moving objects. *Geoinformatica*, 13, 4 (2009), 355-396. DOI= http://dx.doi.org/10.1007/s10707-008-0050-7

[3] Chomicki, J. and Revesz, P. Z. 1999. Constraint-based interoperability of spatiotemporal databases. *Geoinformatica,* 3, 3 (September 1999), 211-243. DOI= http://dx.doi.org/10.1023/A:1009849314891

[4] Garfield, E. 1986. Do Nobel Prize winners write citation classics? *Essays of an Information Scientist*, 9 (1986), 182-187.

[5] Gingras, Y. and Wallace, M. L. 2010. Why it has become more difficult to predict Nobel Prize winners: A bibliometric analysis of nominees and winners of the chemistry and physics prizes (1901-2007). *Scientometrics* 82, 2 (2010), 401-412.

[6] Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A*. 102 (November 2005), 16569-16572. DOI= http://dx.doi.org/10.1073/pnas.0507655102

[7] Kanellakis, P. C., Kuper, G. M. and Revesz, P. Z. 1995. Constraint query languages. *Journal of Computer and System Sciences*, 51, 1 (August 1995) 26-52. DOI= http://dx.doi.org/10.1006/jcss.1995.1051

[8] Li, L. and Revesz, P. Z. 2004. Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems,* 28, 3 (May 2004), 201-227. DOI= http://dx.doi.org/10.1016/S0198-9715(03)00018-8

[9] Ponomarev, I. V., Williams, D. E. , Hackett, C. J. , Schnell, J. D., and Haak, L. L. 2014. Predicting highly cited publications: A method for early detection of candidate breakthroughs, *Technological Forecasting and Social Change*, 81 (January 2014), 49-51. DOI= http://dx.doi.org/10.1016/j.techfore.2012.09.017

[10] Revesz, P. Z. 2010. *Introduction to Databases: From Biological to Spatio-Temporal*, Springer, New York, NY.

[11] Revesz, P. Z. 1997. On the semantics of arbitration. *International Journal of Algebra and Computation*, 7, 2 (April 1997), 133-160. DOI= http://dx.doi.org/10.1142/S0218196797000095

[12] Revesz, P. Z. and Triplet, T. 2011. Temporal data classification using linear classifiers. *Information Systems*, 36, 1, (March 2011), 30–41. DOI= http://dx.doi.org/10.1016/j.is.2010.06.006

[13] Revesz, P. Z. and Triplet, T. 2010. Classification integration and reclassification using constraint databases. *Artificial Intelligence in Medicine*, 49, 2 (June 2010), 79-91. DOI= http://dx.doi.org/10.1016/j.artmed.2010.02.003

[14] Revesz, P. Z. and Wu, S. 2006. Spatiotemporal reasoning about epidemiological data. *Artificial Intelligence in Medicine*, 38, 2 (2006), 157-170. DOI= http://dx.doi.org/10.1016/j.artmed.2006.05.001

[15] Wang, D., Song, C., and Barabási, A. L. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (October 2013), 127-132. DOI= http://dx.doi.org/10.1126/science.1237825